Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 263-266 recibido 26-06-10 revisado 03-07-10 aceptado 06-07-10

Drug-Drug Interaction Detection: A New Approach Based on Maximal Frequent Sequences *†

Detección de Interacciones entre Fármacos: Una nueva aproximación basada en Secuencias Frequentes Maximales

Sandra García-Blasco, Roxana Danger, Paolo Rosso
Natural Language Enginnering Lab. - ELiRF
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
sangarbl@posgrado.upv.es, {rdanger,prosso}@dsic.upv.es

Resumen: En este artículo se presenta un nuevo enfoque para la detección de interacciones entre fármacos. El método propuesto consiste en descubrir patrones automaticamente a través de Secuencias Frecuentes Maximales, y utilizar *pattern matching* para identificar oraciones que contengan interacciones entre fármacos. Las Secuencias Frecuentes Maximales definen secuencias de palabras que son frecuentes en textos y se ha probado en esta investigación que pueden ser un buen método para la detección de interacciones entre fármacos, obteniendo valores prometedores en precisión y cobertura. El método propuesto es independiente del dominio y del lenguaje de los textos.
Palabras clave: Interacciones entre fármacos, extracción de relaciones, secuencias frecuentes maximales

Abstract: In this paper, a new approach for Drug-Drug Interaction detection is presented. The proposed method consists in discovering patterns automatically through Maximal Frequent Sequences extraction and using pattern matching to identify sentences that contain Drug-Drug Interactions. Maximal Frequent Sequences define word sequences that are frequent in texts and it has been proved in this paper to be a good method for DDI detection, obtaining promising results with high values of precision and recall. The method proposed is domain and language independent.
Keywords: Drug-drug Interaction, pattern matching, relation extraction, maximal frequent sequences

1 Introduction

A drug-drug interaction (DDI) occurs when the effects of a drug are modified by the presence of other drugs and its consequences may be very harmful for the patient's health. It is very important that health-care professionals keep their databases up-to-date with respect to new DDI. The growing amount of new medical information makes necessary to find efficient methods to better deal with all this information.

(Segura-Bedmar, 2010) presents two techniques for DDI detection in biomedical texts. The first approximation is a hybrid approach, combining shallow parsing and pattern matching. The patterns used in this technique were described by a pharmacist. With this approach, the authors obtained 48.7% precision and 25.7% recall. The second approach is based in kernel methods, and obtained 55% precision and 84% recall.

In this research work we propose a different approximation for DDI detection in biomedical texts, based in automatically determining the patterns that identify DDI from a training set of documents. Since biomedical texts are written in natural language, a drug drug interaction might be described in so many ways. Our hypothesis holds that there must be some patterns that we can find repeated if we look thought a large amount of biomedical texts, and those patterns will help to identify new drug drug

* The authors would like to thank Isabel Segura-Bedmar and Paloma Martínez for sharing their DrugDDI corpus and Santiago M. Mola for his ideas and technical support. We would also like to thank the reviewers for their valuable comments.
† This work has been partially supported by the MICINN project TEXTENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i) and "Juan de la Cierva" program of the Spanish Ministerio de Ciencia e Innovación.

interactions. The method that we propose is language and domain independent.

This paper is organized as follows. In Section 2 we define *Maximal Frequent Sequences*. Section 3 describes the process followed to identify DDI with *Maximal Frequent Sequences*. In Section 4 we will draw some conclusions.

2 Maximal Frequent Sequences

As presented in (Ahonen-Myka, 2002) a *sequence* is an ordered list of elements, i.e. words. The *frequency* of a sequence is the number of sentences where it appears. A sequence will be β-*frequent* if it is included in β sentences. A *maximal sequence* is a sequence that is not a subsequence of any other. In other words, a maximal sequence shall not be included in any other sequence in the same order. *Maximal Frequent Sequences* (MFS) will be all the sequences that appear in β sentences and that are not subsequences of any other MFS.

In order to make this maximal frequent sequences more flexible, the concept of *gap* is introduced, (García, 2007). The *gap* is the maximum distance that is allowed between two words of a MFS. Following this, if we set the *gap* to 0, the word in the MFS will be adjacent words in the original text. For example, $< w_{i_0}, \ldots, w_{i_n} >$, with $i_j \in 1 \ldots k$, is a maximal frequent sequence of k words, $i_j = i_{j-1} + 1$, $j > 1$, when $gap = 0$, and $i_j \leq i_{j-1} + \eta + 1$, when $gap = \eta$.

MFS have been used for different tasks as measuring text similarities (García-Blasco, 2009) and authorship attribution (Coyotl-Morales et al., 2006).

The algorithm employed to extract MFS is based on the *Apriori Algorithm* (Agrawal and Srikant, 1994), but with the difference that our algorithm takes into account the sequentiality of the elements, allowing gaps between them.

The algorithm takes a collection of sentences and three parameters: $freq_{min}$, gap, $length_{min}$, and is divided in two main steps. First, it extracts all the possible two-word permutations of the set of frequent words, i.e. words that appear at least in $freq_{min}$ sentences. Permutations that are not ordered, respecting the maximum gap allowed, in at least $freq_{min}$ sentences are discarded. Step 2 consists in merging the permutations to form longer sequences. Those that do not fulfill the order and maximum gap conditions are discarded. The algorithm stops when merging candidates is no longer possible. Sequences that are contained in other sequences are discarded.

3 Identifying Drug-Drug Interactions with MFS

Maximal Frequent Sequences will be used to extract the patterns that will allow us to automatically identify drug drug interactions. For each of the Maximal Frequent Sequences extracted we will determine how likely is for that MFS to describe a Drug-Drug Interaction, and then apply it to a test set of biomedical documents to see its performance.

3.1 Corpus

The DrugDDI corpus (Segura-Bedmar, 2010) is a drug-drug interaction corpus annotated with linguistic information, named entities and drug interactions. Drugs are tagged in the corpus, according to their type (clinical drug, antibiotic, etc).

The corpus consists of 579 documents from the DrugBank database, with an average of 10.3 sentences and 5.46 interactions per document.The corpus has been divided into two sets. The first one consists of 446 documents and will be used as the training set. The second set consists of 133 documents and will be our test set.

3.2 Corpus Preprocessing

Taking advantage of the annotations in the corpus, two different preprocessing methods were applied to the original training set. The first one consisted in replacing all the drug names that appeared in the text by their type, i.e. clnd, antb, etc. We will refer to this dataset as *6drugs*. The second preprocessing method consists in replacing all the drug names by the word *#drug#*. We will refer to this dataset as *#drug#*. When we talk about the dataset *norm*, we will refer to the original training set, without any preprocessing.

3.3 Experiments

The objective of this experimentation is to identify drug drug interactions in biomedical texts using *maximal frequent sequences*.

Different sets of MFS were extracted from the training set using different parameters. The algorithm was executed with the three

different versions of the corpus, for $freq_{min}$ in {10,15,20} and *gap* in {0,1,2}.

The MFS detected were rated using a new function that we define, *likeliness*, that is the probability of the MFS of describe a DDI. Likeliness is calculated as:

$$likeliness(MFS_i) = \frac{\text{times } MFS_i \text{ identifies DDI}}{\text{times } MFS_i \text{ appears}}$$

3.4 Results

The algorithm has detected maximal frequent sequences that describe drug-drug interaction. The MFS found have an average length between 4.09 and 4.51 depending on the parameters and the preprocessing of the corpus.

As explained in section 3.3, each MFS has associated a *likeliness* value, that is an indicator of how likely is the MFS to describe a DDI. Figure 1 shows the amount of MFS found for the different corpus, with $freq_{min} = 20$. The bars are also divided according to the likeliness of the MFS.

For example, using the *#drug#* corpus, with req_{min}*=10* and *gap* = 1, the MFS: *('#drug#', 'may', 'the', 'effects', 'of', '#drug#')* was found. This MFS was extracted from sentences like:

- Acetazolamide *may* increase *the effects of* other folic acid antagonists
- Alcohol *may* potentiate *the* side *effects of* bromocriptine mesylate
- Concomitant administration of other sympathomimetic agents *may* potentiate *the* undesirable *effects of* FORADIL

To calculate the performance of the method, the measures of precision, recall and F_1-measure are used. *Precision* is defined as the number of sentences describing DDI retrieved divided by the total number of sentences retrieved, and *Recall* is defined as the number of sentences describing DDI retrieved divided by the total number of existing sentences describing DDI. F_1-measure is the *harmonic mean* of precision and recall.

In Figure 2 the F_1-measure is shown for the different preprocessing and *gap*. With a greater gap, recall grows but it obtains less precision. The threshold established for determining if a MFS describes a DDI, the *likeliness*, plays an important role in the performance of the method. For preprocessing *#drug#* and *6drugs*, the best threshold is in the range $[0.6, 0.7]$. For the normal text, without preprocessing, the best threshold is in the range $[0.1, 0.5]$.

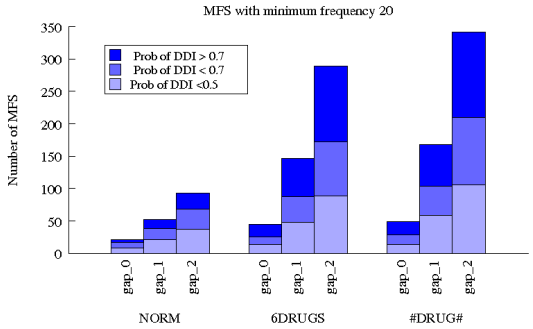


Figure 1: Number of MFS and their likeliness

	Precision	Recall	F_1
baseline	0.5	0.4	0.44
6drugs	0.48	0.93	**0.63**
norm	**0.68**	0.41	0.51
#drug#	0.46	**0.95**	0.62

Table 1: Comparison of Results

Observing the MFS extracted, we can find different types of sequences. Those that have a high value of *likeliness* can be mostly divided in two big groups, those which contain verbs that denote effects, i.e. increase, decrease, enhance, etc., and those which contain 2 or more drugs. Table 2 shows some examples of this two types of MFS found in the documents, their frequency and *likeliness*.

Table 1 gives an overview of the results obtained in the experiments, with *gap=2* and $freq_{min}$*=10*.

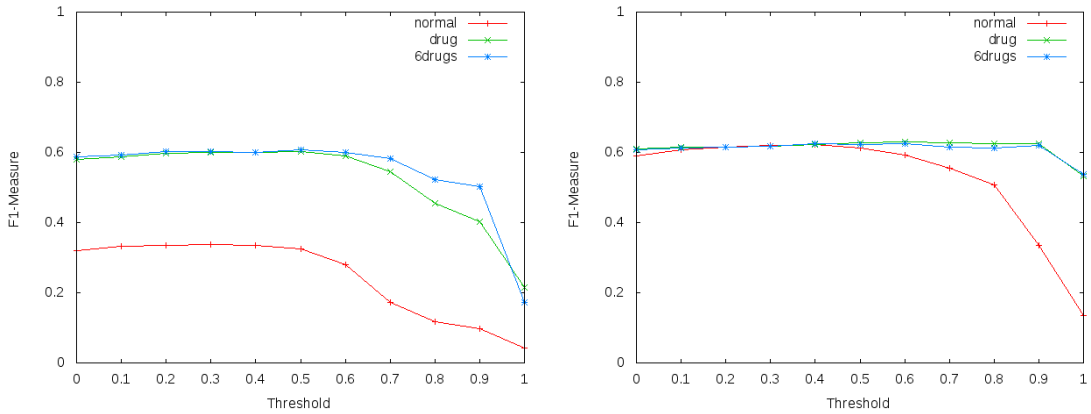
The test set consists of 1151 sentences, with 461 of them describing DDI. As a baseline for this task, the results of a random detector are given. Table 1 contains a relation of the results obtained in this research.[1]

As Table 1 shows, some of the parameters give a very high recall value (95%). DDIs are described by the researchers using a reduced vocabulary and similar sentence structures, i.e. *"Amiodarone should be used with caution*

[1]The results are not directly comparable with those presented in (Segura-Bedmar, 2010) since we calculate precision and recall based on the number of sentences and they do so based on the number of relations. In a sentence several relations might appear. For example, the sentence *"Quinidine and procainamide doses should be reduced when either is administered with amiodarone."* contains two relations: DDI_1*(Quinidine,amiodarone)* and DDI_2*(procainamide,amiodarone)*.

MFS description	Sample	*freq*	*likeliness*
With verbs denoting effects	('#drug#', 'may', 'increase', 'of')	30	0.93
	('may', 'decrease', 'the', 'of')	21	0.90
	('#drug#', 'may', 'enhance', 'the', 'of')	10	1.0
	('#drug#', 'is', 'administered', 'with')	21	0.81
With 2 or more drugs	('#drug#', 'may', 'the', 'effects', '#drug#')	13	1.0
	('#drug#', 'should', 'not', 'be', 'with', '#drug#')	11	1.0
	('#drug#', 'reduce', 'the', 'of', '#drug#')	15	0.93

Table 2: Examples of the MFS extracted



(a) gap=0 (b) gap=2

Figure 2: F_1 for $freq_{min}$=10

in patients receiving propranolol". This allows us to find a set of MFS that retrieve the great majority of the DDIs described. However, the same sentence structures are sometimes used in other contexts, i.e. *"It should be used with caution in patients with diabetes"*. This sentence does not define a DDI, but it does contain a MFS with high likeliness value and it will be labeled as DDI descriptor, decreasing precision.

4 Conclusions

In this paper a new approach to Drug-Drug Interaction detection has been presented. The method presented is domain and language independent, and has been proved to be a good technique for DDI detection.

As further work, we pretend to apply this method to other problems, like Protein-Protein or Protein-Drug Interaction detection. Also, we could enrich the results by adding drug entity identification.

References

Agrawal, Rakesh and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499.

Ahonen-Myka, Helena. 2002. Discovery of frequent word sequences in text. In *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery*, pages 180–189, London, UK. Springer-Verlag.

Coyotl-Morales, Rosa M., Luis Villaseñor-Pineda, Manuel Montes-y Gómez, and Paolo Rosso. 2006. Authorship attribution using word sequences. In *LNCS*, pages 844–853. Springer.

García, René A. 2007. *Algoritmos para el descubrimiento de patrones secuenciales maximales*. Ph.D. thesis, INAOE. Mexico, September.

García-Blasco, Sandra. 2009. *Extracción de secuencias maximales de una colección de textos*. Final degree project, ETSInf, Universidad Politécnica de Valencia, Spain, December.

Segura-Bedmar, Isabel. 2010. *Application of Information Extraction techniques to pharmacological domain: Extracting drug-drug interactions*. Ph.D. thesis, Universidad Carlos III, Madrid, Spain, April.